University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2021

Längsschnittliche Messung numerischer Kompetenzen von Kindergartenkindern: Analysen zur Messinvarianz am Beispiel des Tests TEDI-MATH

Kuratli Geeler, Susanne ; Grob, Urs ; Heinze, Aiso ; Leuchter, Miriam ; Lindmeier, Anke ; Vogt, Franziska ; Moser Opitz, Elisabeth

Abstract: Zusammenfassung. Wird ein Instrument für längsschnittliche Messungen eingesetzt, muss es zusätzlich zu üblichen messtheoretischen Anforderungen die Voraussetzung zeitbezogener Messinvarianz erfüllen. Dies ist für die Erfassung numerischer Kompetenzen im Kindergartenalter aufgrund des hohen Entwicklungstempos herausfordernd. Ziel des vorliegenden Beitrags ist, anhand einer Stichprobe von 894 Kindergartenkindern aus der Schweiz (n = 523) und Deutschland (n = 371) zu untersuchen, ob auf Grundlage des TEDI-MATH-Itempools die Entwicklung numerischer Kompetenzen von Kindergartenkindern reliabel und unverzerrt erfassbar ist. In einer Skalierung mittels des Raschmodells erwies sich das Instrument nach Ausschluss von 17 Items mit zeitbezogenen differenziellen Itemfunktionen (DIF) als überzeitlich reliabel. Eine anschließende CFA zeigte jedoch auf, dass die faktorielle Struktur zwar theoriekonform, über die Zeit jedoch nicht vollständig strukturstabil war. Zudem wies das gekürzte Instrument DIF auf Ebene der Länderteilstichproben auf: Diese Unterschiede könnten durch unterschiedliche Kindergarten-Förderkonzepte in den Ländern bedingt sein, was bereits für das Kindergartenalter die Frage der Kontextabhängigkeit von Leistungsmessungen aufwirft.

DOI: https://doi.org/10.1026/0012-1924/a000262

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-199544
Journal Article
Published Version

The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:
Kuratli Geeler, Susanne; Grob, Urs; Heinze, Aiso; Leuchter, Miriam; Lindmeier, Anke; Vogt, Franziska; Moser Opitz, Elisabeth (2021). Längsschnittliche Messung numerischer Kompetenzen von Kindergartenkindern: Analysen zur Messinvarianz am Beispiel des Tests TEDI-MATH. Diagnostica:1-13.
DOI: https://doi.org/10.1026/0012-1924/a000262

Originalarbeit

Längsschnittliche Messung numerischer Kompetenzen von Kindergartenkindern

Analysen zur Messinvarianz am Beispiel des Tests TEDI-MATH

Susanne Kuratli Geeler[1], Urs Grob[2], Aiso Heinze[3], Miriam Leuchter[4], Anke Lindmeier[3], Franziska Vogt[1] und Elisabeth Moser Opitz[2]

[1]Lehr-Lernforschung, Pädagogische Hochschule St. Gallen, Schweiz
[2]Institut für Erziehungswissenschaft, Universität Zürich, Schweiz
[3]Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik
[4]Universität Koblenz-Landau

Zusammenfassung: Wird ein Instrument für längsschnittliche Messungen eingesetzt, muss es zusätzlich zu üblichen messtheoretischen Anforderungen die Voraussetzung zeitbezogener Messinvarianz erfüllen. Dies ist für die Erfassung numerischer Kompetenzen im Kindergartenalter aufgrund des hohen Entwicklungstempos herausfordernd. Ziel des vorliegenden Beitrags ist, anhand einer Stichprobe von 894 Kindergartenkindern aus der Schweiz (n = 523) und Deutschland (n = 371) zu untersuchen, ob auf Grundlage des TEDI-MATH-Itempools die Entwicklung numerischer Kompetenzen von Kindergartenkindern reliabel und unverzerrt erfassbar ist. In einer Skalierung mittels des Raschmodells erwies sich das Instrument nach Ausschluss von 17 Items mit zeitbezogenen differenziellen Itemfunktionen (DIF) als überzeitlich reliabel. Eine anschließende CFA zeigte jedoch auf, dass die faktorielle Struktur zwar theoriekonform, über die Zeit jedoch nicht vollständig strukturstabil war. Zudem wies das gekürzte Instrument DIF auf Ebene der Länderteilstichproben auf: Diese Unterschiede könnten durch unterschiedliche Kindergarten-Förderkonzepte in den Ländern bedingt sein, was bereits für das Kindergartenalter die Frage der Kontextabhängigkeit von Leistungsmessungen aufwirft.

Schlüsselwörter: Längsschnittmessung, Messinvarianz, Raschanalyse, Kindergarten, numerische Kompetenz

The Longitudinal Assessment of Numerical Competence of Preschoolers: Analyzing Measurement Invariance With the Test TEDI-MATH

Abstract: An instrument used for longitudinal studies must fulfill the criteria of time-related measurement invariance in addition to the general scientific criteria. This is challenging for the assessment of numerical competencies in kindergarten due to the high pace of development at that age. The aim of this analysis was to investigate whether the development of numerical competencies of kindergarten children can be reliably and undistortedly measured with a test based on the TEDI-MATH item pool. The sample consisted of 894 kindergarten children from Switzerland (n = 523) and Germany (n = 371). Based on the Rasch model analysis, the instrument proved to be reliable over time after excluding 17 items with time-related differential item functions (DIF). A subsequent CFA showed, however, that the factorial structure was theoretically consistent but not completely stable over time. In addition, the abridged instrument showed DIF at the level of the country subsamples. These differences could be caused by different kindergarten pedagogies; they also raise the question of to what extent assessments carried out as early as at kindergarten age depend on the context.

Keywords: longitudinal study, measurement invariance, Rasch analysis, kindergarten, numerical competence

In den vergangenen 20 Jahren wurde eine große Anzahl von Studien zu numerischen Vorkenntnissen von Kindern im Kindergartenalter bzw. in der Schuleingangsphase sowie zu geeigneten Fördermaßnahmen durchgeführt. Die Ergebnisse zeigen zum einen, dass numerische Kompetenzen (z. B. Zählen, Zahlen lesen, Anzahlen vergleichen) ein zentraler Prädiktor für die weitere arithmetische Entwicklung in der Grundschule sind (Gallit et al., 2018; Jordan, Kaplan, Locuniak & Ramineni, 2007; Krajewski & Schneider, 2009; Toll, Kroesbergen & van Luit, 2016).

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Zum anderen wurde in mehreren Längsschnittstudien nachgewiesen, dass sich die genannten Kompetenzen bereits im Kindergarten entwickeln und gefördert werden können (Hauser, Vogt, Stebler & Rechsteiner, 2014; Jörns, Schuchardt, Grube & Mähler, 2014; Langhorst, Hildenbrand, Ehlert, Ricken & Fritz, 2013).

Längsschnittuntersuchungen und die damit verbundenen Veränderungsmessungen beinhalten allerdings methodische Herausforderungen, insbesondere wenn es um die Erfassung der Kompetenzentwicklung im Kindergarten- und frühen Schulalter geht. In diesem Alter ist der Entwicklungsstand heterogen und das Entwicklungstempo sehr hoch. Ein einfaches, wenige Testaufgaben umfassendes Instrument kann daher kaum über einen längeren Zeitraum eingesetzt werden. Passt das Instrument zum frühen Testzeitpunkt gut zu den Fähigkeiten der Kinder, ist die Wahrscheinlichkeit hoch, dass es zu einem späteren Zeitpunkt zu einfach ist (Aunio, Heiskari, van Luit & Vuorio, 2015). Würde das Instrument aber ausschließlich auf den späteren Testzeitpunkt hin optimiert, ist zu erwarten, dass die Aufgaben beim ersten Einsatz zu schwierig sind.

Daten sind zudem immer mit Messfehlern behaftet. Bei wiederholter Messung kumulieren sich diese und die Reliabilität der Differenzwerte sinkt (Rost, 2004; Ittel & Merkens, 2006). Es ist deshalb für längsschnittliche Analysen von besonderer Bedeutung, hoch reliable Erhebungsinstrumente einzusetzen, um Messfehler möglichst begrenzt zu halten. Bei der Veränderungsmessung spielt zudem der Faktor Zeit auch in methodischer Hinsicht eine Schlüsselrolle: Damit die längsschnittliche Messung einer Merkmalsdimension als reliabel gelten kann, muss das Instrument u. a. zeitbezogen messinvariant sein, d. h. es darf auf Indikatorebene nicht mit der Zeitvariable interagieren (Putnick & Bornstein, 2016).

Ein möglicher Ausweg besteht darin, für unterschiedliche Testzeitpunkte zwei verschiedene Instrumente einer Testreihe einzusetzen. Dies ist möglich für das Messen der Kompetenzentwicklung vom Kindergarten bis ins erste Schuljahr (Langhorst et al., 2013), jedoch nicht innerhalb des Kindergartenalters, da hier bislang geeignete Tests fehlen. Eine andere Möglichkeit, den genannten Herausforderungen zu begegnen, besteht darin, nur einen Test einzusetzen, der aber ein breites Kompetenzspektrum abdeckt. Ein solcher Test muss sowohl sehr einfache als auch sehr schwierige Items enthalten. Das birgt allerdings die Gefahr eines Boden- oder eines Deckeneffekts. Wäre dies der Fall, erfüllten die an den Enden des Schwierigkeitsspektrums liegenden Items die psychometrischen Anforderungen nicht und trügen unzureichend zur Unterscheidung von Kindern mit unterschiedlicher Kompetenz bei (Döring & Bortz, 2016).

Mathematische Entwicklung im Kindergarten erfassen: Forschungsstand

Die eingangs beschriebenen Herausforderungen der längsschnittlichen Erfassung mathematischer Kompetenzen zeigen sich auch in den vorhandenen Studien, in denen im Kindergarten zwei oder mehrere Male derselbe Test eingesetzt worden ist. In einigen Untersuchungen fehlen die Angaben zur psychometrischen Qualität der längsschnittlich eingesetzten Instrumente (Hauser et al., 2014) oder die Angaben beziehen sich nicht auf die untersuchte Stichprobe (Jörns et al., 2014). Jordan et al. (2007) setzten verschiedene numerische Aufgaben (z. B. Zählen, Vorgänger-Nachfolger bestimmen, einfache Textaufgaben) zu vier Messzeitpunkten während eines Kindergartenjahres ein. Cronbachs Alpha lag dabei immer über .80. Es wurden jedoch keine Analysen zur Messinvarianz durchgeführt. Krajewski und Schneider (2009) haben zu zwei Messzeitpunkten (Abstand 4 Monate) Aufgaben zum Zählen, zum Mengenvergleich und zur Verbindung von Zahlen und Anzahlen eingesetzt. Cronbachs Alpha betrug für das Zählen .72 und .64 für die anderen Aufgaben, wobei nicht klar wird, ob sich diese Werte auf beide Messzeitpunkte beziehen. Aunio et al. (2015) haben eine Skala zum Verständnis von Mengen und Relationen sensu Piaget (Vergleich Mehr-Weniger, Klassifikation, Seriation, Eins-zu-Eins-Zuordnung) sowie zum Zahlwissen und zum Zählen (Zählen, bildgestützte Additionsaufgaben) zu drei Messzeitpunkten im Kindergarten eingesetzt. Testkennwerte werden nicht berichtet. Die Autorinnen und Autoren weisen zudem darauf hin, dass sich ein Deckeneffekt zeigte und Lernfortschritte deshalb nur bei den leistungsschwachen Kindern festgestellt werden konnten. Lyons, Bugden, Zheng, de Jesus und Ansari (2018) setzten drei Aufgabentypen mit je 20 Items zum Zahlen- und Mengenvergleich zweimal während eines Kindergartenjahres ein: Vergleich von zwei Zahlen (α = .83), Vergleich von zwei Punktemengen (α = .70) und Vergleich von Punktemenge und Zahl (α = .69). Für t2 liegen keine Angaben vor. In keiner der Studien kam ein probabilistisches Testmodell zum Einsatz und die verwendeten Skalen wurden auch nicht auf Messinvarianz überprüft. Das steht belastbaren Aussagen zur Struktur und zur Entwicklung der mutmaßlich erfassten Kompetenzen im Wege.

Insgesamt zeigen diese Ausführungen, dass hinsichtlich des Einsatzes von Instrumenten zur Entwicklung der numerischen Kompetenzen im Kindergarten Forschungsbedarf besteht. Dazu sind Tests mit einem großen Itemangebot mit unterschiedlicher Aufgabenschwierigkeit erforderlich und es empfiehlt sich die Abstützung auf die probabilistische Testtheorie.

Eines der wenigen Instrumente, das in einem sehr breiten Altersspektrum eingesetzt werden kann, ist der Test

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

TEDI-MATH (Kaufmann et al., 2009). Es liegen Normen vom vorletzten Kindergartenjahr bis zur dritten Grundschulklasse vor. Der Leistungsheterogenität wird durch eine große Anzahl von Subtests und Items (z. B. in unterschiedlichen Zahlenräumen) Rechnung getragen. Allerdings fehlen Angaben zur dimensionalen Struktur des Tests und es liegen keine Informationen zur auf die Zeit bezogene Messinvarianz vor. Zudem ist die Reliabilität nicht für alle Testzeitpunkte befriedigend. Das Cronbachs Alpha der Kernbatterie für den hier interessierenden Bereich des Kindergartens liegt zwischen .58 und .78. Weiter haben Garrote, Moser Opitz & Ratz (2015) festgestellt, dass bei einzelnen Subtests (z. B. Entscheidung arabische Zahl, Größenvergleich arabische Zahl) die Ratewahrscheinlichkeit mit bis zu 50 % problematisch hoch liegt. Schließlich ist die Durchführung der Kernbatterie trotz einer reduzierten Anzahl von Subtests und Items zeitintensiv und dauert auch im Kindergarten bis zu 45 Minuten.

Forschungsziel

Das Ziel des vorliegenden Beitrags ist erstens, auf Grundlage des Itempools des TEDI-MATH (Kaufmann et al., 2009) ein Testinstrument zu entwickeln, das geeignet ist, die numerische Kompetenz von Kindergartenkindern reliabel zu erfassen und deren Entwicklung über einen Zeitraum von 15 Monaten unverzerrt nachzuzeichnen. Die Datengrundlage[1] bildet eine Stichprobe von Kindergartenkindern (n = 894) aus der Schweiz (n = 523) und aus Deutschland (n = 371) mit drei Testzeitpunkten (mittleres Alter t1 = 5.2 Jahre, t2 = 5.7 Jahre, t3 = 6.5 Jahre).

Gemäß dem aktuellen Forschungstand (Aunio et al., 2015) wird davon ausgegangen, dass die mathematischen Kompetenzen der Kinder vom ersten bis zum dritten Testzeitpunkt insgesamt stark zunehmen. Unklar hingegen ist, ob sich die Kompetenzen in allen Bereichen numerischer Kompetenz gleich entwickeln oder ob bezogen auf einzelne Kompetenzaspekte unterschiedliche Entwicklungsverläufe stattfinden. Unter Nutzung der Möglichkeiten des Rasch-Modells zur Diagnose diffenzieller Itemfunktionen (DIF; Rasch, 1960; Strobl, 2015) und durch konfirmatorische Faktorenanalysen wird daher das Instrument zweitens auf zeitbezogene Messinvarianz hin untersucht.

Schließlich interessiert drittens, ob sich in der deutschen und der schweizerischen Stichprobe unterschiedliche Teilaspekte numerischer Kompetenz differenziell entwickeln, da sich der Kindergarten in der Schweiz und die Kindertagesstätte in Deutschland bezüglich der Konzepte mathematischer Frühförderung und der Förderpraxis unterscheiden (Kuratli Geeler, 2019). Während die Kindertagesstätten in Deutschland ein außerfamiliäres (kostenpflichtiges) Betreuungsangebot mit einem Erziehungsauftrag darstellen, hat der Kindergarten in der Schweiz einen Bildungsauftrag und gehört in nahezu allen Kantonen zur obligatorischen Schule. Im Lehrplan der Schweiz werden spezifisch mathematische Kompetenzen beschrieben, an denen im Kindergarten verbindlich gearbeitet werden soll (Deutschschweizer Erziehungsdirektorenkonferenz, 2014). Die Bildungspläne in Deutschland basieren auf den Empfehlungen der Jugendministerkonferenz / Kultusministerkonferenz (JMK / KMK, 2004), stellen aber keine verbindlichen Zielvorgaben sondern Anregungen dar (Diskowski, 2009). Mathematik ist dabei als Thema vorgesehen (JMK / KMK, 2004), wird aber in den verschiedenen Bundesländern sehr unterschiedlich umgesetzt. Es kann daher angenommen werden, dass in den Kindergärten der Schweiz eine intensivere mathematische Förderung stattfindet. Diese kann sich auf die Entwicklung der Lösungsraten bei bestimmten inhaltlichen Anforderungen auswirken, die in den schweizerischen Kindergärten speziell gefördert werden. Das vierte Ziel des Beitrags besteht entsprechend darin, Unterschiede in den Itemcharakteristika nach Land und Zeit zu untersuchen.

Veränderungsmessung mit dem Rasch-Modell

Das Rasch-Modell (Rasch, 1960) gehört zur probabilistischen Testtheorie bzw. der Item-Response-Theory (IRT) und gilt als speziell geeignet, latente Eigenschaften bzw. Kompetenzen und deren Entwicklung über die Zeit zu erfassen (Strobl, 2015). Beim eindimensionalen Rasch-Modell wird die Wahrscheinlichkeit für die Lösung von Testaufgaben in Abhängigkeit von der geschätzten Kompetenzausprägung der untersuchten Person modelliert (Rost, 2004; Döring & Bortz, 2016).

Dabei geben itembezogene Passungswerte (fit indicators) Aufschluss darüber, wie gut einzelne Aufgaben bzw. Indikatoren ins Modell passen. Dies erlaubt den Ausschluss ungeeigneter Items. Zentral sind dabei die gewichteten Mean-Square-Abweichungen (Infit-MNSQ). Ein Wert von 1 bedeutet eine ideale Passung in das Messmodell, Werte über 1 indizieren für das jeweilige Item einen Mangel an Vorhersagbarkeit der Antworten aus dem Personenschätzwert, Werte unter 1 eine zu hohe Vorhersag-

[1] Die Daten der Untersuchung wurden im Rahmen eines vom Schweizerischen Nationalfonds (Kennzeichen 100019L-156680) und von der Deutschen Forschungsgemeinschaft (Kennzeichen LI 2616/1-1, HE 4561/8-1, LE 3327/2-1) geförderten Forschungsprojektes erhoben.

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

barkeit. Für den praktischen Umgang mit Infit-MNSQ-Werten werden unterschiedliche Toleranzgrenzen vorgeschlagen. Boone, Staver und Yale (2014) erachten einen Bereich von 0.50 bis 1.50 als tolerierbar. Wilson (2005) empfiehlt als Grenzwerte 0.75 und 1.33. Bond und Fox (2015) befürworten für Multiple-Choice-Tests Grenzwerte von 0.80 und 1.20.

Für den Umgang mit der Zeit einer längsschnittlichen Datenstruktur bieten sich in der Item-Response-Theorie zwei Verfahren an: Das Schätzen eines mehrdimensionalen Rasch-Modells (Rasch, 1961) mit jeweils einer Merkmalsdimension pro Testzeitpunkt oder die Behandlung der wiederholten Messungen in Form von „virtuellen Personen" (Hartig & Kühnbach, 2006). Letzteres hat gegenüber dem erstgenannten Verfahren den Nachteil, dass Abhängigkeiten zwischen den Messwerten einer Person nicht explizit berücksichtigt werden. Beide Vorgehensweisen führen jedoch bei der Verwendung von ML-basierten Schätzern (MLE und WLE[2]) zu annähernd gleichen Personenparameterwerten (Hartig & Kühnbach, 2006.).

Zu den wichtigsten Vorteilen des Rasch-Modells zählt die Eigenschaft, für Aufgaben des Typs richtig / falsch ein explizites Messmodell darzustellen, dessen Gültigkeit für verschiedene Subgruppen und für wiederholte Messungen über die Zeit getestet werden kann. Nachweisliche Messinvarianz stellt die erforderliche Grundlage dar für Aussagen zu Gruppenvergleichen und Veränderungen über die Zeit. Zur Überprüfung von auf Subgruppen und auf die Zeit bezogener Messinvarianz können DIF-Analysen auf Itemebene vorgenommen werden (Schwab & Helm, 2015). Diese nehmen die Lage der Schwierigkeitsparameter der Items für (apriori-) Gruppen bzw. zu den Testzeitpunkten in den Blick. Die relative Schwierigkeit aller Aufgaben in Logit-Werten sollte sich über die Gruppen oder Testzeitpunkte hinweg im Idealfall vollständig parallel darstellen. Wahrscheinlichkeitstheoretisch testbare Abweichungen hiervon bringen die Verletzung der Annahme vollständig gleicher Eigenschaften der Items und damit des Tests über die Gruppen bzw. die Zeit hinweg zum Ausdruck. Da neben der wahrscheinlichkeitsbezogenen Perspektive auch die Frage des Grades der Abweichung relevant ist, stützt sich beispielsweise der Educational Testing Service (ETS) auf ein kombiniertes Klassifikationssystem der (Nicht-) Einhaltung von Messinvarianz im Kontext des Rasch-Modells (Longford, Holland, & Thayer, 1993; Zwick, Thayer & Levis, 1999). Dieses kombiniert eine wahrscheinlichkeitstheoretische mit einer effektstärkebezogenen Perspektive. Gemäß dessen pragmatischer Stufenlogik können Parameterdifferenzen bis maximal 0.426 (entsprechend einem Mantel-Haenszel-Delta von 1.0) bei nicht gegebener Signifikanz ($p \geq .05$) als vernachlässigbar gelten. Ist $p < .05$ und die Differenz der Parameter kleiner als 0.638 (was einem Delta von 1.5 entspricht), gilt der Grad an DIF als mittel. Ist die Differenz hingegen 0.638 oder größer (bei $p < .05$), gilt das Ausmaß an DIF als groß.

Prüfung von Messinvarianz mit der konfirmatorischen Faktorenanalyse

Zur Prüfung der zeitbezogenen Messinvarianz können alternativ zu den DIF-Analysen im Raschmodell konfirmatorische Faktorenanalysen durchgeführt werden. Dabei werden die binären Aufgabenlösungen mittels linearer Regression zur Schätzung vorgegebener Facetten der Personfähigkeit verwendet. Messtheoretisch ist dieses Vorgehen nicht ideal. Es bietet aber den Vorteil, dass die Stufe der gültigen zeitbezogenen Messinvarianz bestimmt werden kann (Meredith, 1993; Brown, 2006). Bei der am wenigsten restriktiven Stufe der konfiguralen Messinvarianz besteht einzig eine äquivalente Faktorenstruktur. Bei der metrischen bzw. schwachen faktoriellen Invarianz werden zusätzlich zu beiden Zeitpunkten die Faktorladungen als äquivalent angenommen, bei der skalaren bzw. starken faktoriellen Invarianz zusätzlich die Intercepts der manifesten Variablen und bei der strikten Invarianz auch noch die Messfehlerinvarianzen. Unterschiedliche Analysen setzen jeweils andere Stufen der Invarianz voraus. Für den Vergleich korrelativer Beziehungen über zwei Testzeitpunkte hinweg muss mindestens metrische Invarianz gegeben sein, für den Vergleich von latenten Mittelwerten mindestens skalare.

Methode

Messinstrumente

Die numerischen Kompetenzen wurden mit einer angepassten Version des TEDI-MATH (Kaufmann et al., 2009) erhoben. In der Originalversion umfasst der TEDI-MATH für das zweite Kindergartenjahr acht Subtests in der Kernbatterie (88 Items) und zusätzlich neun Subtests für die Gesamtbatterie (59 Items). Für die vorliegende Untersuchung wurden mehrere Anpassungen vorgenommen. Zwei Subtests (Entscheidung zur Korrektheit von Zahl-

[2] MLE = Maximum Likelihood Estimator, WLE = Weighted Likelihood Estimator nach Warm (1989), die im Vergleich zu MLE lageabhängig leicht zum Mittelwert hin korrigiert sind

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Tabelle 1. Demographische Angaben der Untersuchungsstichprobe

	Deutschland	Schweiz	Gesamt
Anzahl Klassen	73	67	140
Anzahl Kinder	371	523	894
Geschlecht w / m	179 / 192	258 / 265	437 / 457
Alter t1 in Jahren *M* (*SD*)	5.12 (0.36)	5.25 (0.37)	5.20 (0.37)

wörtern wie „zweizehn"; Textaufgaben) wurden ausgeschlossen, da die erfolgreiche Lösung dieser Items stark von den sprachlichen Kompetenzen der Kinder abhängt. Um Deckeneffekte zu vermeiden, wurden zudem drei Subtests (Transkodieren, Ordnen nach numerischer Größe, Subtraktion) aufgenommen, die in der Originalversion für ältere Kinder vorgesehen sind.

Beim Größenvergleich von arabischen Zahlen wurde zur Reduktion der Ratewahrscheinlichkeit jedes Zahlenpaar zweimal in jeweils umgekehrter Reihenfolge präsentiert (z. B. 2 vs. 6 und später 6 vs. 2) und nur dann als richtig bewertet, wenn beide Antworten korrekt waren. In der Aufgabe zur additiven Zerlegung (Subtest 10) wurden die Zahlen durch konkrete Mengen (Abbildungen von Schafen) ersetzt. Das Lösen der Subtraktionsaufgaben mit Objektabbildungen (Subtest 11) wird in der Originalversion mit Bildern unterstützt, in denen Minuend und Subtrahend nicht eindeutig abgebildet sind (z. B. fünf verschiedenfarbige Bälle zur Aufgabe 5 – 2). Die Bilder der Subtraktionsaufgaben wurden deshalb durch Abbildungen ersetzt, in denen Minuend und Subtrahend eindeutig sichtbar sind (z. B. fünf Ballone, zwei Ballone fliegen weg). In der TEDI-MATH-Originalversion werden die Aufgaben zur unvollständigen Addition und Subtraktion (Subtest 13 und 14) für ältere Kinder auf der formal-symbolischen Ebene dargeboten. Diese Aufgaben wurden für die Kindergartenstufe mit Plättchen veranschaulicht („Hier siehst du vier Plättchen. Wie viele Plättchen muss ich dazu legen, damit es sechs Plättchen sind?").

Der Subtest 7 (Ordnen von Zahlen) enthielt nur ein Item. Hier wurde ein zusätzliches Item gebildet (Einordnen einer Zahlenkarte in eine Zahlenreihe). Schließlich wurden zwei neue Subtests aufgenommen, die gemäß Krajewski (2008) „Meilensteine" der numerischen Entwicklung sind und die in der Originalversion fehlen: Zum einen wurde ein Item entwickelt (Subtest 8), bei dem einer gegebenen Menge eine Zahl zugeordnet werden muss. Zum anderen (Subtest 15) wurde mit fünf Items das Verständnis von Beziehungen zwischen zwei Zahlen überprüft („Hier siehst du die Zahl fünf. Welche Zahl ist um zwei größer als fünf?").

Insgesamt umfasste das revidierte Instrument 95 Items, davon wurden 81 Items zu allen drei Testzeitpunkten eingesetzt (Ankeritems). Da im vorliegenden Beitrag die zeitbezogene Messinvarianz im Vordergrund steht, wurden nur diese 81 Items einer Skalierung nach dem Rasch-Modell mit DIF-Analysen unterzogen (Überblick über Subtests, Anzahl Items und Anpassungen siehe ESM 1). Die Testdurchführung dauerte 20 bis 25 Minuten und war somit deutlich kürzer als die Durchführung der Kernbatterie in der Originalversion. Des Weiteren wurden das Alter (in Jahren), das Geschlecht, die Zugehörigkeit zur Lerngruppe und damit auch zum Land erfasst.

Stichprobe

Analysiert wurden Daten von 894 Kindergartenkindern aus 67 Kindergärten in der Schweiz (n = 523) und aus 73 Kindertagesstätten in Deutschland (n = 371). Tabelle 1 gibt einen Überblick der demographischen Angaben der Untersuchungsstichprobe. Es wurden nur Kinder in die Stichprobe einbezogen, bei denen die numerische Kompetenzen zu allen drei Testzeitpunkten gemessen werden konnten. Der Drop-out von t1 zu t3 betrug n = 251. Es handelt sich dabei um zufällige Ausfälle einzelner Kinder (Wegzug, Fehlen wegen Krankheit) oder ganzer Kindergartenklassen (Mutterschutz oder Stellenwechsel von Lehrpersonen). Die mathematischen Kompetenzen der Teilstichprobe von Kindern mit zu allen drei Zeitpunkten vollständigen Testwerten (n = 894, M = 31.99, SD = 14.43) und der Teilstichprobe mit fehlenden Werten beim zweiten / und oder dritten Testzeitpunkt (n = 251, M = 30.22, SD = 15.20) unterschieden sich zum ersten Messzeitpunkt nicht signifikant voneinander (t (1 143) = 1.702, p = .09).

Analysen

Die Rasch-Analysen wurden mit Hilfe der Statistiksoftware R (R Core Team [Version 3.5.3], 2019) und des Pakets TAM von Robitzsch, Kiefer und Wu (2017) durchgeführt, wobei der Mittelwert der Itemschwierigkeiten auf 0 festgesetzt wurde. Die Personenparameter sind wie im Standardmodell über eine 50-prozentige Lösungswahrscheinlichkeit an die Itemschwierigkeit gekoppelt. Zunächst wurden die Items nach Testzeitpunkt getrennt skaliert und bei allen Items wurde der gewichtete Item-Fit

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

(Infit-MNSQ) bestimmt, der Aufschluss gibt, wie gut jedes Item ins Modell passt. Gemäß den Empfehlungen von Wilson (2005) wurden Items mit einem Infit-MNSQ zwischen 0.75 und 1.33 toleriert bzw. außerhalb liegende Items ausgeschlossen.

In einem zweiten Schritt wurde mittels DIF-Analysen nach Land überprüft, ob die Schwierigkeitsparameter der Items in jedem Land zum ersten Testzeitpunkt die gleiche relative Lage einnahmen. Items mit nach ETS-Kategoriensystem „großer" differenzieller Itemfunktion wurden ausgeschlossen.

Im dritten Schritt wurden die verbliebenen Items auf zeitbezogene Messinvarianz geprüft. Da zwischen dem ersten und dritten Testzeitpunkt der Kompetenzzuwachs am größten ist, konnte davon ausgegangen werden, dass die Items für diesen längsten Zeitraum am anfälligsten für Messvarianz waren. Deshalb wurden DIF-Analysen zwischen dem ersten und dritten Testzeitpunkt durchgeführt, indem die Daten pro Testzeitpunkt als Teilstichproben (virtuelle Personen) betrachtet wurden. So konnte die Lage der Schwierigkeitsparameter zu den beiden Testzeitpunkten miteinander verglichen und Items mit problematisch hoch ausgeprägter zeitbezogener DIF („groß" nach ETS) Schritt für Schritt iterativ ausgeschlossen werden. Das resultierende Instrument umfasste ausschließlich Items mit DIF unterhalb der ETS-Stufe „groß", d. h. alle Parameterdifferenzen waren entweder nicht signifikant von 0 verschieden ($p \geq .05$) oder, falls signifikant, geringer ausgeprägt als 0.638 Logits.

Um zu untersuchen, ob zwischen der Stichprobe in Deutschland und der Stichprobe in der Schweiz Unterschiede bezüglich zeitbezogener Messinvarianz bestehen, wurden zusätzlich DIF-Analysen zwischen dem ersten und dritten Testzeitpunkt getrennt nach Land durchgeführt. Für die Interpretation und Diskussion der länderspezifischen zeitbezogenen Messvarianz interessierten Items mit hoher DIF.

Zur Beantwortung der Frage, ob die faktorielle Struktur der Kompetenzaspekte, die sich in für gleichartige Aufgaben konvergierenden Testleistungen spiegeln, über die Zeit stabil ist, wurden mit 54 Items, die keine oder nur begrenzte DIF aufwiesen, Faktorenanalysen durchgeführt. Zunächst wurde die faktorielle Struktur auf Grundlage eines Datensatzes, in dem die Messungen des späteren Testzeitpunkts als virtuelle Personen behandelt und mit den Messungen zu t1 zusammengeführt wurden, mittels einer explorativen Faktorenanalyse (EFA) bestimmt. Um für das modifizierte Instrument die jeweilige Stufe der faktoriellen Invarianz zwischen t1 und t3 zu bestimmen, wurden in Mplus 8.3 drei konfirmatorische Faktorenanalysen (CFA) geschätzt. Dazu wurden die Daten längsschnittlich behandelt, d. h. die identischen Indikatoren für die beiden Testzeitpunkte (t1 und t3) wurden über separate Variablen repräsentiert. Es wurde ein Gesamtmodell mit je einer strukturidentischen CFA pro Testzeitpunkt geschätzt. Die Fehlerterme der identischen Indikatoren wurden paarweise über die Zeit korreliert. Zwischen den identischen Faktoren von t1 und t3 wurden paarweise Korrelationen zugelassen, ebenso unter allen Faktoren jeweils innerhalb eines Testzeitpunkts, nicht jedoch über die Testzeitpunkte hinweg.

Die abschließenden längsschnittlichen Auswertungen zum Lernzuwachs der Kinder wurden auf Grundlage des revidierten Instruments vorgenommen. Mittels der Variante mit virtuellen Personen wurden für die drei Testzeitpunkte Personenparameter des Typs WLE (Warms Weighted Likelihood Estimate) berechnet (Warm, 1989), die mittels Varianzanalyse mit Messwiederholung statistisch analysiert wurden. Dies erfolgte mit Hilfe der Statistiksoftware R (R Core Team [Version 3.5.3], 2019) und des Pakets EZ von Lawrence (2016).

Ergebnisse

Analysen getrennt nach Testzeitpunkt

Die WLE-Reliabilität für die zu den drei Testzeitpunkten separat geschätzten Rasch-Modelle (je 81 identische Items) war insgesamt hoch und betrug für t1 Rel = .95, für t2 Rel = .96 und für t3 Rel = .95. Bei t1 lag der Infit-MNSQ für alle Items im tolerierbaren Bereich (0.75 bis 1.33). Bei t2 wiesen zwei Items einen leichten Underfit auf. Betroffen waren ein Item aus dem Subtest 4 zum Größenvergleich mit MNSQ = 1.39 und ein Item zum approximativen Punktevergleich aus Subtest 16 mit MNSQ = 1.38. Die Analysen bei t3 ergaben wiederum für zwei Items einen leichten Underfit. Der MNSQ für ein Item aus dem Subtest Abzählen (2) betrug 1.43, für ein Item aus dem Subtest Größenvergleich (4) 1.34. Die genannten vier Items mit problematischem Fitwert wurden ausgeschlossen. Die weiteren Analysen zur Messinvarianz bezogen sich in der Folge auf 77 zu jedem Testzeitpunkt identische Items.

DIF Analysen zu Land und Testzeitpunkt

DIF nach Land bei t1. Im revidierten Instrument mit 77 Items waren bei t1 sechs Items von großer DIF nach Land betroffen: ein Item zum Zählen in 10er-Schritten aus Subtest 1 (Parameterdifferenz $\Delta\beta = 1.054$), eines zum Abzählen aus Subtest 2 ($\Delta\beta = 0.640$), drei Items zum Größenvergleich aus Subtest 4 ($\Delta\beta = 0.784 / 1.062 / 0.952$) und ein Item zur Addition aus Subtest 12 ($\Delta\beta = 0.812$).

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

Tabelle 2. Übersicht Items mit großer zeitbezogener DIF von t1 zu t3 für die Gesamtstichprobe und getrennt für die Stichproben aus der Schweiz und aus Deutschland

			Parameterdifferenz		
Subtest (Nummer)	Variable	Aufgabe	Gesamtstichprobe	Schweiz	Deutschland
Zahlwortreihe (1)	zaeh_12b	vorwärts von 7 bis 15	0.506	***0.682***	0.166
	zaeh_13a	rückwärts von 7	***0.950***	***1.072***	0.626
	zaeh_13b	rückwärts von 15	0.608	***0.720***	0.470
	zaeh_14a	2er-Schritte vorwärts	***1.380***	***1.426***	0.542
Abzählen (2)	abz_21a	5 Objekte abzählen	-0.556	***-1.062***	-0.118
	abz_21c	5 Objekte abgedeckt, Nachfrage 2: Wie viele?	***-0.742***	-0.376	***-1.030***
	abz_22a	9 Objekte abzählen	***-1.034***	***-1.342***	-0.006
	abz_22c	9 Objekte zugedeckt, Nachfrage 2: Wie viele?	***-0.730***	-0.458	***-0.908***
	abz_23c	12 Objekte zugedeckt, Nachfrage 2: Wie viele?	-0.536	-0.070	***-0.784***
	abz_24	Gleiche Anzahl legen	***-0.650***	-0.586	***-0.700***
Größenvergleich (4)	gva_41	23 vs. 14	***-0.668***	***-0.902***	0.088
Zahlen lesen (5)	trz_51	3	***0.958***	***0.902***	***0.776***
	trz_52	6	***0.730***	***0.696***	***0.914***
	trz_53	8	0.506	0.432	***0.704***
	trz_513	105	-0.420	***-0.922***	-0.032
Ordnen nach num. Größe: Zahlen (7)	ordz_71	Ordnen nach Größe (Zahlen)	0.408	***0.736***	0.308
	ordz_72	Zahl richtig einordnen	0.406	***0.660***	0.366
Numerische Inklusion (9)	inkl_91	Im Briefumschlag sind 6 Plättchen. Kannst du mir 8 daraus geben?	***-2.442***	***-2.626***	***-1.488***
Rechnen mit Objektabbildungen (11)	obj_112	5 und 3 Bleistifte	***-0.824***	***-1.074***	-0.448
	obj_116	9 weg 2 Büchsen	-0.502	***-0.686***	-0.160
Addition (12)	add_121	2 + 2	***0.890***	***0.934***	**0.680**
	add_123	6 + 3	0.634	***0.700***	0.286
	add_125	6 + 6	***0.758***	***0.834***	0.250
	add_126	9 + 4	0.448	***0.682***	-0.118
	add_128	20 + 8	***0.756***	***0.701***	0.100
Unvollständige Subtraktion (14)	uvs_141	3 Plättchen auf der Unterlage, wie viele wegnehmen, damit es 1 sind?	***-0.646***	***-0.834***	-0.246
	uvs_142	8 Plättchen auf der Unterlage, wie viele wegnehmen, damit es 5 sind?	-0.578	***-0.662***	-0.364
Approx. Punktevergleich (16)	sgv_162	7 und 2	***-0.734***	-0.304	***-1.404***
	sgv_163	7 und 12	***-0.678***	***-0.804***	-0.612

Anmerkungen: Die markierten Items mit großer DIF (Differential Item Functioning), gemäß Kriterium nach ETS, waren mit $p < .05$ signifikant.

Diese sechs Items wurden für die nachfolgenden Analysen zur zeitbezogenen Messinvarianz ausgeschlossen.

Zeitbezogene DIF von t1 zu t3 in der Gesamtstichprobe. Die Ergebnisse der DIF-Analysen zur Messinvarianz zwischen dem ersten und dritten Testzeitpunkt (basierend auf dem revidierten Instrument mit 71 Items) sind in Tabelle 2 dargestellt. Eine positive Parameterdifferenz bedeutet, dass das Item zum dritten Testzeitpunkt verglichen mit der mittleren Entwicklung aller Items einfacher zu lösen war. Eine negative Parameterdifferenz bedeutet umgekehrt, dass das Item zum späteren Zeitpunkt im Vergleich zur mittleren Entwicklung aller Items schwieriger zu lösen war bzw. in geringerem Ausmaß einfacher wurde.

Die Ergebnisse zeigen, dass in der Gesamtstichprobe insgesamt 17 Items gemäß ETS-Kriterien (siehe Veränderungsmessung mit dem Rasch-Modell) ein großes Maß an zeitbezogener DIF aufwiesen (siehe Tabelle 2). Betroffen waren insbesondere Items aus den Subtests Zahlwortreihe (1) und Abzählen (2). Hier waren insgesamt 6 von 14 Items von zeitbezogener DIF betroffen. Auch in den Subtests Addition (12) und Approximativer Punktevergleich (16) wiesen mehrere Items DIF auf. Zudem gab es vereinzelt Items mit DIF aus weiteren Subtests (siehe Tabelle 2). In den Subtests Entscheidung arabische Zahl (3), Ordnen nach numerischer Größe (6 und 7), Zahlen-Größen-Zuordnung (8), Additive Zerlegung (10),

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

Tabelle 3. Passungswerte der CFA für die 54 ausgewählten Aufgaben über zwei Testzeitpunkte für konfigurale und metrische Invarianz

	Chi^2 ΔChi^2	*df* Δ *df*	*p*	RMSEA	SRMR	CFI	TLI	BIC
Modell 1 Konfigurale Invarianz	13285.9	5429	.000	0.040	0.134	0.827	0.815	56733.8
Modell 2 Metrische / schwache faktorielle Invarianz	13870.2 584.3	5469 40	.000 .000	0.041	0.138	0.814	0.804	57046.2
Modell 3 Skalare / starke faktorielle Invarinaz	15515.7 2229.8	5520 91	.000 .000	0.045	0.159	0.779	0.769	58345.2

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Unvollständige Addition (13) und Beziehungen zwischen Zahlen (15) waren dagegen keine Items von Messvarianz über die Zeit betroffen. Alle von großer DIF betroffenen Items wurden für den letzten Analyseschritt ausgeschlossen und im Instrument verblieben schließlich 54 Items.

Zeitbezogene DIF von t1 zu t3 für die Schweizer und die Deutsche Stichprobe. Bezüglich der getrennt für die beiden Länderstichproben untersuchten zeitbezogenen Messinvarianz, zeigten sich deutliche Unterschiede (Tabelle 2). Die Stichprobe der Schweiz war mit insgesamt 23 Items im Vergleich zur Stichprobe in Deutschland mit insgesamt 10 Items deutlich stärker von Messvarianz betroffen. In den Schweizer Daten wurde eine starke zeitbezogene DIF bei Items in den Subtests Zahlwortreihe (1), Zahlen nach der Größe ordnen (7), Addition (12) und Unvollständige Subtraktion (14) festgestellt. In den Daten aus Deutschland waren es vor allem Items aus dem Subtest Abzählen (2) und einige Items aus dem Subtest Zahlen lesen (5), die von DIF betroffen waren. Die so festgestellte DIF nach Land x Zeit wurde als informativ, aber als nicht korrigierbar betrachtet. Es erfolgte deshalb kein Ausschluss weiterer Items. Eine Übersicht zu den Parameterdifferenzen aller Items zwischen dem ersten und dritten Testzeitpunkt wird im Elektronischen Supplement 2 (ESM 2) zur Verfügung gestellt.

Faktorielle Struktur der 54 geeigneten Items des angepassten Instruments

Die Analyse der faktoriellen Struktur erfolgte bezogen auf die 54 Items, die nicht von zeitbezogener DIF betroffen waren. Die in einem ersten Schritt mit den zusammengeführten Daten von t1 und t3 vorgenommene EFA ergab unter Rückgriff auf den Eigenwertverlauf (SCREE-Kriterium) und die Plausibilität der Struktur 14 Faktoren, wobei drei nur durch je eine Aufgabe bestimmt waren. Die 11 Faktoren mit mehreren Indikatoren entsprachen mit zwei Ausnahmen (Ordnen nach numerischer Größe Zahlen / Anzahlen und unvollständige Addition / Subtraktion) den Subtests des TEDI-MATH. Beide Ausnahmen bestehen allerdings nur darin, dass sich je zwei Teilaspekte der übergeordneten Aufgabentypen faktorenanalytisch nicht als trennbar erwiesen und stattdessen auf einen gemeinsamen Faktor luden.[3] Im Anschluss wurde diese faktorielle Struktur (11 Faktoren und 3 Einzelitems) auf ein längsschnittliches Modell mit je einer strukturidentischen CFA pro Testzeitpunkt angewendet. Die Ladungen dieser CFA sind in ESM 3 dokumentiert. Während sich darin die durch die EFA bestimmte Struktur des angepassten TEDI-MATH – mit den oben genannten zwei Ausnahmen – für beide Zeitpunkte durch in der Regel genügend hohe Ladungen bestätigte, wiesen pro Zeitpunkt fünf Indikatoren eine problematisch niedrige Ladung (< .50) auf (siehe ESM 3).

Bereits die CFA zur konfiguralen Invarianz (Tabelle 3, Modell 1) wies mit Ausnahme des RMSEA problematische Fitwerte auf. Der Chi^2-Wert war signifikant, was auf Misfit ausserhalb des Zufallsstreubereichs hindeutete, der SRMR lag über dem als sinnvolle Obergrenze geltenden Wert von 0.08, die relativen Fitmasse CFI und TLI blieben klar unterhalb der Schwelle von .95 für eine vertretbare Passung des Modells auf die Daten (Hu & Bentler, 1999). Es liegt somit weder konfigurale noch metrische oder skalare Invarianz vor. Die Verschlechterung der Passungswerte der CFAs zur metrischen und skalaren Invarianz (Tabelle 3 Modelle 2 und 3) verdeutlichen den Befund der Nichtpassung des Modells auf die Daten zusätzlich.

Längsschnittliche Analysen

Für die längsschnittlichen Analysen wurden basierend auf dem revidierten Instrument mit 54 Items mittels *virtueller Personen* die WLE-Personenfähigkeitsparameter zu jedem der drei Testzeitpunkte innerhalb eines einzigen Modells geschätzt. Die WLE-Reliabilität betrug 0.95. Die

[3] Siehe ESM 3, faktorielle Struktur der CFA basierend auf der Annahme konfiguraler Invarianz.

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

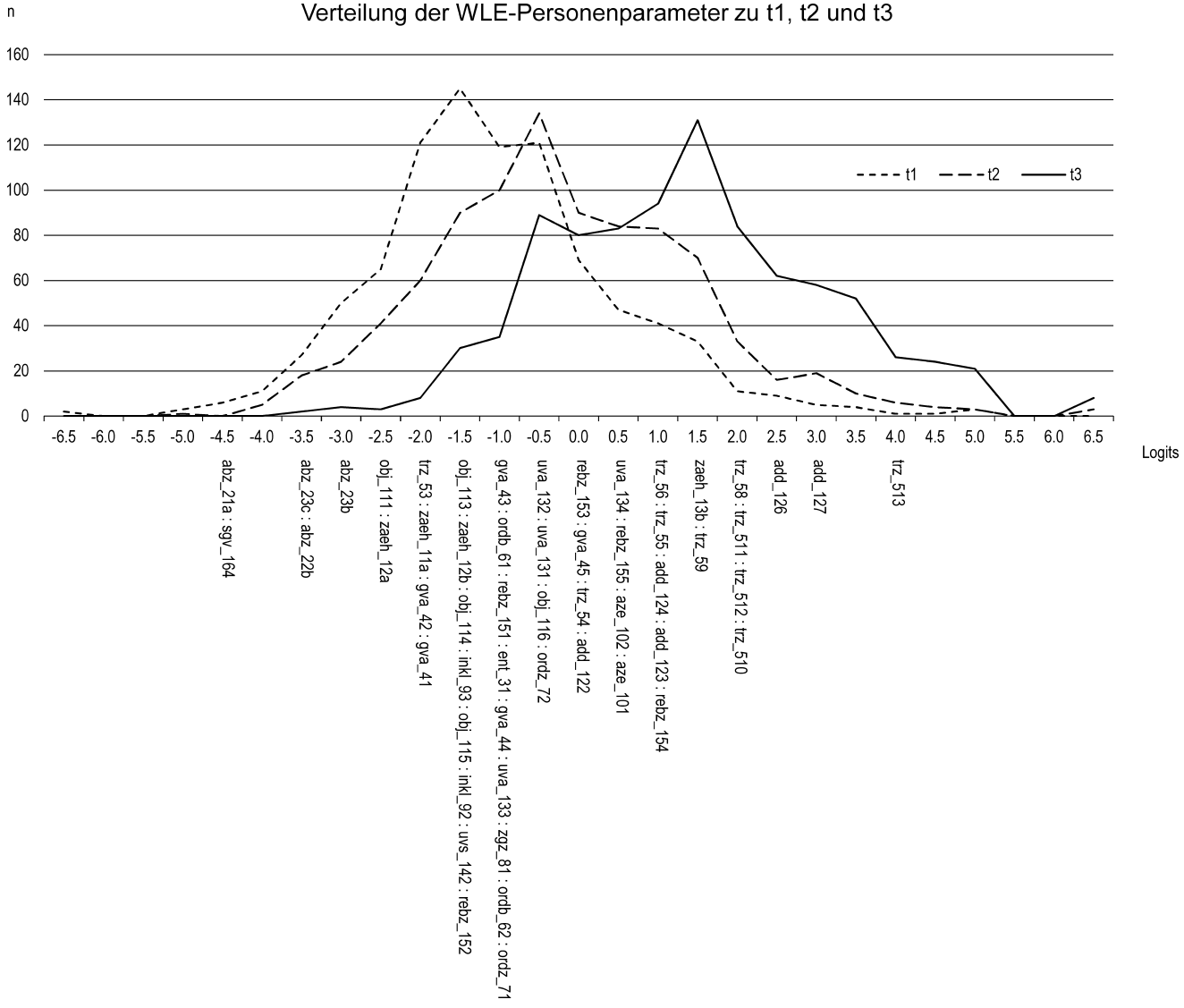

Abbildung 1. Verteilung der WLE-Personenparameter zu t1, t2 und t3. (Der Tabelle 2 bzw. dem ESM 2 kann entnommen werden, welche Aufgaben hinter den Kürzeln stehen).

Personenparameter lagen beim ersten Testzeitpunkt zwischen -6.30 und 5.03 mit einem Mittelwert von -1.06 (*SD* = 1.50), beim zweiten zwischen -5.05 und 6.29 mit einem Mittelwert von -.23 (*SD* = 1.66) und beim dritten Testzeitpunkt zwischen -3.57 und 6.29 mit einem Mittelwert von 1.32 (*SD* = 1.70). Die deskriptiven Testauswertungen zeigten zudem, dass kein Boden- und Deckeneffekt vorhanden war. Es gab bei keinem Testzeitpunkt Kinder ganz ohne richtig gelöste Aufgaben und lediglich acht Kinder von 894 (1%) erreichten beim dritten Testzeitpunkt die maximale Punktzahl.

In Abbildung 1 sind die drei Verteilungen (t1 bis t3) der Kinder (Anzahl auf y-Achse) nach Fähigkeit bzw. nach Schwierigkeit der Items (x-Achse) festgehalten. Die Personen- und die Itemparameter wurden zu Bereichen von jeweils 0.5 Logits zusammengefasst. Die Kurven über die drei Testzeitpunkte haben sich erwartungsgemäß von links nach rechts verschoben. Darin zeigt sich die Zunahme der Personenfähigkeit bzw. der Kompetenzerwerb der Kinder. Diese Zunahme ist statistisch signifikant, wie sich mit einer Varianzanalyse mit Messwiederholung zeigen lässt (F [2,1786] = 2 593.57, $p < .001$, eta^2 = .74, f = 1.70). In den mathematischen Kompetenzen der Kindergartenkinder über die drei Testzeitpunkte werden 74 % der Varianz durch den Faktor Zeit erklärt.

Die Korrelation der Personenfähigkeitsparameter zwischen der ursprünglichen Testversion mit allen Items (81 Items) und der Version nach Ausschluss der kritischen Items (54 Items) betrug zu den drei Testzeitpunkten zwischen 0.97 und 0.99.

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Diskussion

Das Ziel vorliegender Untersuchung war, die mathematische Entwicklung von Kindergartenkindern über 15 Monate auf der Basis der Personen- und Itemparameter aufzuzeigen. Zudem wurde mittels Rasch-Analysen (Rasch, 1960; Strobl, 2015) und konfirmatorischen Faktorenanalysen überprüft, ob ein Itempool aus dem Test TEDI-MATH (Kaufmann et al., 2009) das Kriterium der zeitbezogenen Messinvarianz erfüllt. Des Weiteren interessierte, ob sich bezüglich spezifischer Themen eine differenzielle Entwicklung zeigt und ob es in Bezug auf die Messinvarianz Unterschiede zwischen der Stichprobe in Deutschland und der Stichprobe der Schweiz gibt.

Psychometrische Qualität des angepassten Instruments

Wie dargestellt, wurden bei einigen Items des TEDI-MATH von Kaufmann et al. (2009) Anpassungen vorgenommen, da sich in anderen Untersuchungen Schwachstellen insbesondere hinsichtlich der Ratewahrscheinlichkeit gezeigt hatten (Garrote et al., 2015). Zudem wurden im Testmaterial einige Darstellungen angepasst. Insgesamt kann davon ausgegangen werden, dass die adaptierten Items ins Instrument passen. Durch den Einbezug von Items, die in der Originalversion für ältere Kinder vorgesehen waren, konnte ein Deckeneffekt vermieden werden. Auch ein Bodeneffekt war nicht vorhanden. Damit ist eine wesentliche Voraussetzung der Eignung des Instruments für eine längsschnittliche Leistungsmessung gegeben.

Vier Items wiesen in den nach Testzeitpunkt getrennten Analysen ungünstige Infit-MNSQ-Werte auf. Zwei dieser Items waren sehr schwierig (Größenvergleich), zwei Items (approximativer Mengenvergleich, Abzählen) waren hingegen sehr einfach. Es wird angenommen, dass diese Lage der Schwierigkeitsparameter zu den ungünstigen Fit-Werten beigetragen hat. Sechs Items wiesen bei t1 Messvarianz bezüglich Land auf und 17 Items waren in der Gesamtstichprobe von zeitbezogener Messvarianz betroffen. Die Analysen zur längsschnittlichen Kompetenzentwicklung der Kinder über die drei Testzeitpunkt zeigten, dass nach Ausschluss der problematischen Items die WLE-Reliabililität der verbliebenen 54 Items mit Rel = .95 sehr gut und deutlich höher war als Cronbachs Alpha in der – deutlich kleineren – deutschen Normierungsstichprobe des TEDI-MATH. Die hohen Korrelationen der WLEs mit und ohne Ausschluss von unbefriedigenden Items zu allen drei Testzeitpunkten (≥ .97) weisen darauf hin, dass dasselbe Konstrukt gemessen wurde. Zudem sind immer noch Items aus allen Subtests des Instrumentes vorhanden. Eine inhaltliche Bedeutungsverschiebung der Kompetenzdimension kann deshalb weitgehend ausgeschlossen werden, zugleich relativieren diese Ergebnisse eine mögliche Verzerrung der Leistungsmessung durch von DIF betroffene Items. Aus testökonomischer Sicht ist das ein wichtiger Aspekt, da sich die Dauer der Aufgabenbearbeitung (ca. 20 bis 25 Minuten) nochmals verkürzt und noch ca. einen Drittel der Zeit beansprucht, die für die Durchführung der Kernbatterie in der Originalversion des TEDI-MATH vorgesehen ist.

Zeitbezogene Messinvarianz und differenzielle Entwicklungen

Bezüglich der zeitbezogenen Messinvarianz waren in der Gesamtstichprobe Items aus verschiedenen Subtests von DIF betroffen. Es konnte keine einheitliche, auf einzelne Kompetenzbereiche konzentrierte differenzielle Entwicklung festgestellt werden. Werden die Ergebnisse getrennt für die Stichproben der beiden Länder betrachtet, so zeigen sich dagegen deutliche Unterschiede, was auf eine differenzielle Entwicklung in den beiden Ländern schließen lässt.

In der Stichprobe der Schweiz wurde erstens im Vergleich zur deutschen Stichprobe bei mehr als doppelt so vielen Items eine zu große zeitbezogene Messvarianz festgestellt. Zweitens waren häufig mehrere Items aus einem Subtest betroffen, was auf eine differenzielle Entwicklung einzelner Bereiche hinweist. In der deutschen Stichprobe war dies nicht der Fall. In der schweizerischen Stichprobe gab es bei vier Items zum verbalen Zählen DIF. Das bestätigt die Ergebnisse von Aunio et al. (2015), die in ihrer Längsschnittuntersuchung im Kindergarten bei den Zählaufgaben ausgeprägte Lernfortschritte vorfanden, die zu einem problematischen Deckeneffekt führten. Auch bezüglich des Ordnens von Zahlen, des Rechnens mit Objektabbildungen sowie der Addition wurde in der schweizerischen Teilstichprobe ein höher ausgeprägtes Maß an DIF festgestellt. Bei all diesen Themen handelt es sich um eher prozedurales Wissen, bei dem davon ausgegangen werden kann, dass es sich nicht spontan entwickelt, sondern durch Anleitung erworben und durch Üben gefestigt wird (z. B. das Kennen der Zahlenreihenfolge, das Rückwärtszählen, das Zählen in 2er-Schritten oder das Addieren).

In beiden Länder-Stichproben wiesen mehrere Items zum Lesen von Zahlen DIF auf, zudem auch Items zum Abzählen von Objekten. In der Stichprobe aus Deutschland war das etwas häufiger der Fall als in der Schweiz. Es handelt sich hierbei – im Gegensatz zu den vorher diskutierten Aufgaben – um numerische Kompetenzen,

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

die auch in alltagsbezogenen Lerngelegenheiten erworben werden können, wenn z. B. eine Zahl auf der Fernbedienung abgelesen wird oder bei einem Spiel die Kinder abgezählt werden.

Die Ergebnisse deuten somit darauf hin, dass sich in der schweizerischen Stichprobe im Unterschied zur deutschen Stichprobe bestimmte numerische Kompetenzen, die im Test erfasst wurden, stärker entwickelt haben als andere. Dieser Unterschied könnte curricular bedingt sein, da in der Schweiz verbindliche Lernziele im Fach Mathematik formuliert sind (Deutschschweizer Erziehungsdirektorenkonferenz, 2014). In Deutschland wird zudem eher eine alltagsintegrierte mathematische Förderung fokussiert, während in der Schweiz konkrete numerische Kompetenzen formuliert werden, die verbindlich zu erreichen sind (z. B. im Zahlenraum bis 10 von jeder möglichen Zahl aus vor- und rückwärts zählen, Aussagen zu Anzahlen und Zahlpositionen machen oder Anzahlen verschieden darstellen).

Insgesamt kann davon ausgegangen werden, dass in der deutschen Stichprobe aufgrund der größeren Varianz in den kontextuellen Rahmenbedingungen weniger einheitliche, spezifische Förderung in einzelnen mathematischen Bereichen stattfindet und sich darum die Rangreihenfolge der Itemparameter vom ersten zum dritten Testzeitpunkt weniger stark verändert hat als in der schweizerischen Stichprobe.

Während mittels des Rasch-Modells belegt wurde, dass mit dem auf 54 Aufgaben gekürzten Instrument eine globale Dimension numerischer Kompetenz überzeitlich reliabel gemessen werden kann, muss aus den CFAs geschlossen werden, dass die faktorenanalytisch bestimmten Kompetenzfacetten über die Zeitspanne zwischen t1 und t3 nicht vollständig strukturstabil sind. Allerdings ist relativierend festzuhalten, dass ein durch von 0 abweichende Nebenladungen bedingter Misfit bei einem Messmodell dieser Komplexität (zwei Teil-CFAs mit je 11 bzw. 14 Faktoren für je 54 Indikatoren) und der gegebenen hohen thematischen Homogenität (alle Aufgaben beziehen sich auf Teilaspekte numerischer Kompetenz) – wenig überrascht (Kenny & McCoach, 2003). Der Umstand, dass die Struktur der Subtests des gekürzten TEDI-MATH in der CFA zu konfiguraler Invarianz mit zwar mäßigem Fit, aber nur wenigen strukturellen Schwachpunkten abgebildet werden konnte, kann als erstaunlich bezeichnet werden. Dennoch gilt es festzuhalten, dass sich numerische Kompetenz erhoben mit dem TEDI-Math nicht in überzeitlich vollständig stabile Subdimensionen aufschließen ließ. Die für die Lösung der jeweiligen Aufgaben erforderlichen Teilfähigkeiten scheinen sich in der betrachteten Altersphase nicht völlig parallel zu entwickeln.

Folgerungen für längsschnittliche Messungen mathematischer Kompetenzen im Kindergarten

Im Abschnitt zum Forschungsstand wurde aufgezeigt, dass in den bisher vorliegenden längsschnittlichen Untersuchungen im Kindergarten (Aunio et al., 2015; Hauser et al., 2014; Jörns et al., 2017; Jordan et al., 2007; Krajewski & Schneider, 2009; Lyons et al, 2018) Angaben zur Messinvarianz der eingesetzten Instrumente fehlen. Die Analysen haben erstens gezeigt, dass für die längsschnittliche Leistungsmessung im Kindergarten ein breites Itemangebot, das sowohl im unteren als auch im oberen Leistungsbereich differenziert, notwendig ist. Der TEDI-MATH erfüllt dieses Kriterium, wenn zu der für den Kindergarten vorgesehenen Kernbatterie zusätzliche, für ältere Kinder vorgesehene Items dazu genommen werden. Auf diese Weise gab es keinen Boden- und Deckeneffekt und nur wenige Items zeigten ungenügende Item-Fit-Werte. Die numerischen Leistungen einer sehr heterogenen Gruppe konnten reliabel gemessen und Aussagen über den Lernzuwachs gemacht werden.

Limitationen

Als Limitation der Studie ist festzuhalten, dass sich die Analysen auf eine in Lerngruppen und Kindergärten getestete, nicht-repräsentative Stichprobe von Kindern beziehen, deren pädagogische Fachkräfte sich freiwillig am Forschungsprojekt beteiligt haben. Es ist nicht auszuschließen, dass die Stichprobe deswegen eine positive Selektion darstellt.

Ausblick

Die Ergebnisse der zeitbezogenen DIF-Analysen zeigen, dass das differenzielle Funktionieren von Items als bedeutsame Quelle von Erkenntnis zu disparallelen Entwicklungsprozessen unter bestimmten Bedingungen betrachtet werden sollte. Von DIF betroffene Items geben Auskunft über den Einfluss von potentiell unterschiedlichen Fördermaßnahmen in Bezug auf verschiedene pädagogische Kontexte. Allerdings verweisen die Resultate der CFAs darauf, dass bezüglich der Binnenstruktur numerischem Kompetenz im Kindergartenalter noch nicht der wünschbare Stand an Wissen erreicht ist. Das zeigt sich auch darin, dass in anderen Instrumenten (Krajewski 2018; Ricken, Fritz & Balzer, 2013) die faktorielle Struktur bis anhin nicht überprüft worden ist.

Es wurde sichtbar, dass zwischen den Subdimensionen numerischer Kompetenzen Überschneidungen bestehen.

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Die sowohl in den DIF-Analysen wie den CFAs erkennbaren disparallelen Entwicklungen lassen interdependente Dynamiken bzw. logische Abhängigkeiten von Teilkompetenzen vermuten, die in zukünftigen Forschungsarbeiten vertiefend untersucht werden sollten. Aufgrund der Komplexität numerischer Kompetenz erschiene zum besseren Verständnis der Entwicklungsprozesse die Fokussierung auf eine begrenzte Auswahl von Aufgabentypen sinnvoll, auf die bezogen die Entwicklungsdynamik und die wechselseitige Beeinflussung von Teilkompetenzen beispielsweise mittels aufeinander bezogener Wachstumskurvenmodelle nachgezeichnet werden könnten.

Mit Blick auf das Ziel, numerische Kompetenz im Kindergartenalter möglichst objektiv und beispielsweise im Hinblick auf Vergleiche zwischen Ländern und Institutionen fair zu erfassen, verweisen die Befunde dieses Beitrags auf Grenzen, da sich zeigte, dass institutionelle Kontexte Einfluss auf die Entwicklung numerischer Kompetenzen nehmen können (Garrote & Moser Opitz, 2020). Zwar könnten in der Folge Kompetenzbereiche, deren Entwicklung sich offenbar in Interaktion mit externen Einflüssen differenziell entwickeln, aus Instrumenten entfernt werden. Wenn diese Dimensionen jedoch theoretisch für die globale Kompetenzdimension zentral sind, wäre das eine ungute Strategie, die zu nicht verantwortbaren thematischen Verengungen führte. Stattdessen empfiehlt sich, solche Kontextfaktoren zukünftig als Bestandteil des Hintergrundmodells explizit in die Messmodelle aufzunehmen. Dies gilt nicht nur für die hier fokussierte Leistungsmessung im Kontext des Kindergartens sondern auch für andere vergleichende Zugänge, in denen Interaktionen mit Kontexten, etwa in Form unterschiedlicher oder unterschiedlich zeitlich sequenzierter schulischer Curricula bestehen.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000262

ESM 1. Übersicht Mathematiktest. Die Tabelle zeigt eine Übersicht über Subtests, Anzahl Items und Apassungen gegenüber dem TEDI-MATH (Kaufmann et al., 2009).

ESM 2. Übersicht zeitbezogene Messinvarianz. Die Tabelle zeigt die Parameterdifferenzen zwischen dem ersten und dritten Testzeitpunkt aller Items der Gesamtstichprobe und getrennt nach Ländern, wobei die Items mit großer DIF markiert sind.

ESM 3. Übersicht CFA konfigurative Invarianz. Die Tabelle zeigt die Ladungsstruktur und Passungswerte einer CFA für die 54 ausgewählten Aufgaben über zwei Testzeitpunkte unter Annahme konfiguraler Invarianz.

Literatur

Aunio, P., Heiskari, P., Luit, J. E. van & Vuorio, J.-M. (2015). The development of early numeracy skills in kindergarten in low-, average- and high-performance groups. *Journal of Early Childhood Research*, *13*(1), 3 – 16. https://doi.org/10.1177/1476718X14538722

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). *Rasch analysis in the human sciences.* Dordrecht, NL: Springer.

Brown, T. A. (2006). *Confirmatory factor analysis.* New York, NY: Guilford.

Deutschschweizer Erziehungsdirektoren-Konferenz (D-EDK) (2014). *Lehrplan 21.* Verfügbar unter: https://www.lehrplan.ch/

Diskowski, D. (2009). Bildungspläne für Kindertagesstätten: Ein neues und doch unbegriffenes Steuerungsinstrument. In H. G. Roßbach & H. P. Blossfeld (Hrsg.), *Frühpädagogische Förderung in Institutionen* (S. 47 – 61). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91452-7_4

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin: Springer. https://doi.org/10.1007/978-3-642-41089-5

Gallit, F., Wyschkon, A., Poltz, N., Moraske, S., Kucian, K., Aster, M. von et al. (2018). Henne oder Ei: Reziprozität mathematischer Vorläufer und Vorhersage des Rechnens. *Lernen und Lernstörungen, 7*(2), 81 – 92. https://doi.org/10.1024/2235-0977/a000205

Garrote, A. & Moser Opitz, E. (2020).Erfassung von mathematischen Kompetenzen im Vorschulalter: Erprobung des Tests MARKO-D an einer Schweizer Stichprobe und Überprüfung des Instruments für Messwiederholungen. *Diagnostica* [Vorab-Onlinepublikation], https://doi.org/10.1026/0012-1924/a000258.

Garrote, A., Moser Opitz, E. & Ratz, Ch. (2015). Mathematische Kompetenzen von Schülerinnen und Schülern mit dem Förderschwerpunkt geistige Entwicklung: Eine Querschnittstudie. *Empirische Sonderpädagogik, 7*(1), 24 – 40. Verfügbar unter: https://www.pedocs.de/frontdoor.php?source_opus=10280

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values" in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft* (S. 27 – 44). https://doi.org/10.1007/978-3-531-90502-0_3

Hauser, B., Vogt, F., Stebler, R. & Rechsteiner, K. (2014). Förderung früher mathematischer Kompetenzen: Spielintegriert oder trainingsbasiert. *Frühe Bildung, 3*, 139 – 145. https://doi.org/10.1026/2191-9186/a000144

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1 – 55. https://doi.org/10.1080/10705519909540118

Ittel, A. & Merkens, H. (2006). *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft.* Wiesbaden: Verlag für Sozialwissenschaften.

Jordan, N. C., Kaplan, D., Locuniak, M. N. & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36 – 46. https://doi.org/10.1111/j.1540-5826.2007.00229.x

Jörns, C., Schuchardt, K., Grube, D. & Mähler, C. (2014). Spielorientierte Förderung numerischer Kompetenzen im Vorschulalter und deren Eignung zur Prävention von Rechenschwierigkeiten. *Empirische Sonderpädagogik, 6*, 243 – 259. Verfügbar unter: https://www.pedocs.de/frontdoor.php?source_opus=9933

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1026/0012-1924/a000262 - Wednesday, January 27, 2021 3:39:56 AM - IP Address:188.60.169.102

Jugendministerkonferenz/Kultusministerkonferenz (JMK/KMK) (2004). *Gemeinsamer Rahmen der Länder für die frühe Bildung in Kindertageseinrichtungen* (Beschluss vom 13./14. 05. 2004). Verfügbar unter: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_06_03-Fruehe-Bildung-Kindertageseinrichtungen.pdf

Kaufmann, L., Nuerk, H.-C., Graf, M., Krinzinger, H., Delazer, M. & Willmes, K. (2009). *TEDI-MATH: Test zur Erfassung numerisch-rechnerischer Fertigkeiten vom Kindergarten bis zur 3. Klasse.* Bern: Huber.

Kenny, D. A. & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*, 333 – 351. https://doi.org/10.1207/S15328007SEM1003_1

Krajewski, K. (2018). *MBK 0: Test mathematischer Basiskompetenzen im Kindergartenalter.* Göttingen: Hogrefe.

Krajewski, K. (2008). *Vorhersage von Rechenschwäche in der Grundschule* (2., korr. Aufl.). Hamburg: Dr. Kovač.

Krajewski, K. & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction, 19*, 513 – 526. https://doi.org/10.1016/j.learninstruc.2008.10.002

Kuratli Geeler, S. (2019). *Mathematische Kompetenzen von Kindergartenkindern: Überprüfung eines Testinstrumentes und Analyse von Unterschieden in der numerischen Leistungsentwicklung* (Unveröffentlichte Dissertation). Universität Zürich. https://www.zora.uzh.ch/id/eprint/171088/

Langhorst, P., Hildenbrand, C., Ehlert, A., Ricken, G. & Fritz, A. (2013). Mathematische Bildung im Kindergarten: Evaluation des Förderprogramms „Mina und der Maulwurf" und die Betrachtung von Fortbildungsvarianten. In M. Hasselhorn, A. Heinze & W. Schneider (Hrsg.), *Diagnostik mathematischer Kompetenzen* (S. 113 – 134). Göttingen: Hogrefe.

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments.* Retrieved from https://cran.r-project.org/web/packages/ez/index.html

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171 – 196). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Lyons, I. M., Bugden, S., Zheng, S., De Jesus, S. & Ansari, D. (2018). Symbolic number skills predict growth in nonsymbolic number skills in kindergarteners. *Developmental Psychology*, *54*, 440 – 457. https://doi.org/10.1037/dev0000445

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525 – 543. https://psycnet.apa.org/doi/10.1007/BF02294825

Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71 – 90. https://doi.org/10.1016/j.dr.2016.06.004

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Kopenhagen: The Danish Institute for Education Research.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology.* Berkeley, CA: University of California Press.

R Core Team. (2019). *The R Project for statistical computing.* Verfügbar unter: https://www.r-project.org/

Ricken, G., Fritz, A. & Balzer, L. (2013). *MARKO-D: Mathematik- und Rechenkonzepte im Vorschulalter – Diagnose.* Göttingen: Hogrefe.

Robitzsch, A., Kiefer, T. & Wu, M. (2017). *Package TAM, Test analysis modules.* Verfügbar unter: htttps://CRAN.R-proicet.org/package=TAM

Rost, J. (2004): *Lehrbuch Testtheorie- Testkonstruktion* (2., überarb. und erw. Aufl.). Bern: Hans Huber.

Schwab, S. & Helm, Ch. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik 7*, 175 – 193. Verfügbar unter: http://nbn-resolving.de/urn:nbn:de:0111-pedocs-113801

Strobl, C. (2015). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* (3., erw. Aufl., Bd. 2 Sozialwissenschaftliche Forschungsmethoden). München: Rainer Hampp.

Toll, S. W. M., Kroesbergen, E. H. & Luit, J. E. H. van (2016). Visual working memory and number sense: Testing the double deficit hypothesis in mathematics. *British Journal of Educational Psychology, 86*, 429 – 445. https://doi.org/10.1111/bjep.12116

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427 – 450. https://doi.org/10.1007/BF02294627

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R., Thayer, D. T. & Lewis, C. (1999). An empirical bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement*, *36*(1), 1 – 28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x

Historie
Onlineveröffentlichung: 22. 01. 2021

Dr. Susanne Kuratli Geeler
Lehr-Lernforschung
Pädagogische Hochschule St. Gallen
Notkerstrasse 27
9000 St. Gallen
Schweiz
susanne.kuratli@phsg.ch

© 2021 Hogrefe Verlag Veröffentlicht als Hogrefe OpenMind-Artikel unter der Lizenz CC BY 4.0 (http://creativecommons.org/licenses/by/4.0)